A new optimization method of the geometric distance in an automatic recognition system for bird vocalisations

M. Jinnai[a], N. Boucher[b], M. Fukumi[c] and H. Taylor[d]

[a]Kagawa National College of Technology, 355 Chokushi-cho, 761-8058 Takamatsu, Japan
[b]SoundID, PO Box 649 Maleny, 4552 Queensland, Australia
[c]University of Tokushima, 2-1 Minami-josanjima, 770-8506 Tokushima, Japan
[d]Wissenschaftskolleg zu Berlin Institute for Advanced Study, Wallotstrabe 19, 14193 Berlin, Germany
jinnai@t.kagawa-nct.ac.jp

We have been developing an automatic recognition system for bird vocalisations. Many biologists have been using the early 32 bit version of our system, and we have been working on a 64 bit version. The software segments a waveform of the bird vocalisation from a three-hour continuous recording and extracts the sound spectrum pattern from the waveform using the LPC spectrum analysis. Next, the software compares the sound spectrum pattern (the input pattern) with the standard pattern (that was extracted in advance) using a similarity scale. We use a new similarity scale called the "Geometric Distance". The Geometric Distance is more accurate than the conventional similarities in the noisy environment. In the 64 bit version, the software matches an input pattern with the 40,000 elements of the standard patterns per second and per processor, and it is 2.8 times faster than the conventional cosine similarity. In this paper, we introduce an automatic segmentation method of bird vocalisations and a new optimization method of the Geometric Distance. The new optimization method offers improvements of an order of magnitude over the conventional Geometric Distance.

1 Introduction

The introduction of the ornithological spectrograph in the 1950s changed the way birdsong was measured, quantified, and interpreted by biologists. The spectrograph was correlated with objective knowledge and quickly superseded aural assessment, which was considered more subjective. Given the reliance on visual information in birdsong research, the accuracy of the sound spectrograph is critical. However, the spectrograph remained largely unchanged and unchallenged. With this in mind, we have updated an automatic recognition system for the bird vocalisations over the past nine years [1,2,3]. As of early 2012, we have commercial software that: (1) segments a waveform of the bird vocalisation automatically from a three-hour continuous recording, (2) extracts the spectrum patterns from the bird vocalisation, and (3) matches the spectrum patterns using a similarity scale and recognizes the bird vocalisation.

The spectrum patterns are usually produced using the computationally efficient FFT, which generates much subtle detail that is largely redundant in recognition. Furthermore, because the FFT is ideal for long steady-state signals, it produces artefacts when applied to short bird vocalisations. The LPC (Linear Predictive Coefficient) is much more suitable for such transient signals and does not generate artefacts. Hence we have adopted the LPC, which, despite its computational complexity, is not an issue with today's high-speed computers.

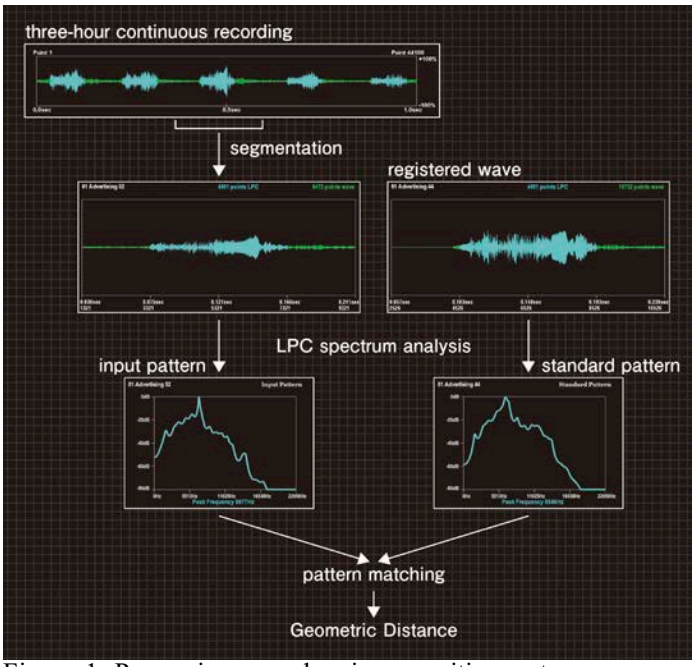

Figure 1: Processing procedure in recognition system.

The similarity scale works as follows: for vocalisations for which a researcher would recognize two patterns as similar to each other, the computer software outputs a small value, and for vocalisations for which a researcher would recognize the two patterns as dissimilar, the computer software then outputs a large value. In conventional sound recognition, the similarity scales known as the Euclidean distance and cosine similarity are widely used to measure likeness. Conventional similarity scales compare the patterns using one-to-one mapping. The result of the one-to-one mapping is that the distance metric is highly sensitive to noise, and the distance metric changes in a staircase pattern when a difference occurs between peaks of the standard and input patterns. As an improvement, we have developed a new similarity scale called the "Geometric Distance (GD)" [4,5]. The GD is more accurate than the conventional similarity scales in the noisy environment. Furthermore, the GD is 2.8 times faster than the conventional cosine similarity.

We expected that the LPC and GD would result in faster and better recognition than that which can be achieved by a human expert, and the 32 bit software has already proven to be comparable to or better than a human expert. We have now moved on to a 64 bit version that is 4 times faster than the 32 bit version of our system. In this paper, we introduce an automatic segmentation method of bird vocalisations and describe a GD algorithm and its new optimization method.

2 Automatic recognition system

Figure 1 shows the processing procedure to recognize bird vocalisations. First, the software segments a waveform of the vocalisation from a three-hour continuous recording. Second, the software extracts the sound spectrum pattern (frequency-power) from the waveform using the LPC spectrum analysis. Third, the software compares the sound spectrum pattern (the input pattern) with the standard pattern (that was registered in advance) using a new similarity scale called the GD and recognizes the bird vocalisation. Section 3 describes the automatic segmentation method of the bird vocalisation. Section 4 describes the LPC spectrum of the bird sounds. Sections 5 and 6 describe pattern matching using the similarity scales.

3 Segmentation of bird vocalisations

Eq. (1) and Figure 2 show the method for calculating an energy curve E_k. We suppose that $w_i\,(i=1,2,\ldots,N)$ is the amplitude of a continuous recording waveform for example from a microphone, where N is the number of the recording

wave data and M is the range of energy calculation. In Figure 2, the calculated energy curve is shown by the yellow line. Note that we calculate the energy curve using $|w_i|$ instead of w_i^2 in order to reduce the processing overhead.

$$E_k = \sum_{i=k-M}^{k+M} |w_i| \qquad (k=1+M, 2+M, \cdots, N-M) \qquad (1)$$

In the actual software, the energy curve E_k can be calculated by the following recurrence formula using long integers E_k and w_i in order to reduce the processing overhead.

$$E_k = E_{k-1} - |w_{k-M-1}| + |w_{k+M}| \qquad (2)$$
$$(k=2+M, 3+M, \cdots, N-M) \qquad E_{1+M} = \sum_{i=1}^{1+2M} |w_i|$$

Figure 3 shows the method for segmenting the bird vocalisation from a three-hour continuous recording using the energy curve E_k. As shown in Figure 3, we set the threshold value arbitrarily in advance. Here, if ($E_j \leq$ threshold and threshold $< E_{j+1}$) and ($E_l \geq$ threshold and threshold $> E_{l+1}$), then we find the position k that corresponds to the maximum value E_k within the range of j to l. Next, we calculate the LPC spectrum using the waveform of the range of $k-M$ to $k+M$ shown by blue line in Figure 3.

The amplitude of the microphone output signal for field recordings of bird vocalisations is highly variable. For example the recording could be of a bird in flight. In this instance, we need to periodically normalize the energy curve for each period of time. Figure 4 shows the method for normalizing the energy curve (Auto Gain Control) with yellow and red lines. In Figure 4, we have set the AGC with a 1.0 second frame width and a 0.7 second frame period. Following the blue line of the waveform shown in Figure 4, we find that the bird vocalisations can be segmented accurately even if the amplitude of the microphone output signal is reduced.

4 The LPC spectrum of bird sounds

The bottom diagrams of Figure 1 show the spectra (frequency-power) extracted from the vocalisations of the critically endangered Coxen's Fig-Parrot (*Cyclopsitta diophthalma coxeni*). These spectra have been calculated using the method of Linear Predictive Coefficient (LPC). We have set the analysis conditions of the bird vocalisation with a 44.1kHz sampling frequency, 16 bit quantization, 90.7 msec frame width, 44 order LPC, 1Hz to 22050Hz frequency range, 86Hz frequency resolution, and 0dB to −80dB logarithmic power spectrum.

From the spectra shown in Figure 1, it is evident that the peak frequencies are 6977Hz and 6546Hz respectively, and the two patterns are somewhat similar to each other. Moreover, as a result of comparing these two bird vocalisations aurally, we have confirmed that these two vocalisations are somewhat similar to each other.

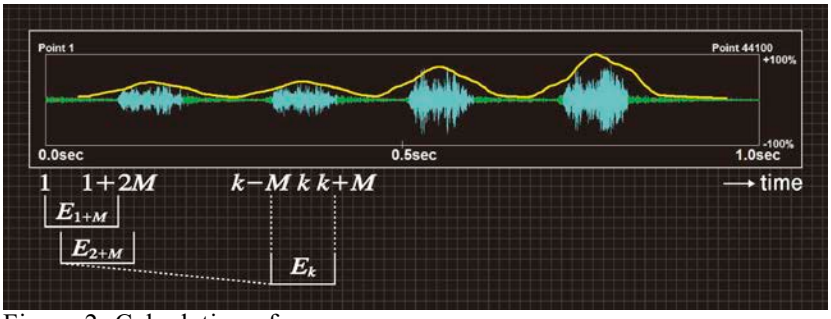

Figure 2: Calculation of energy curve.

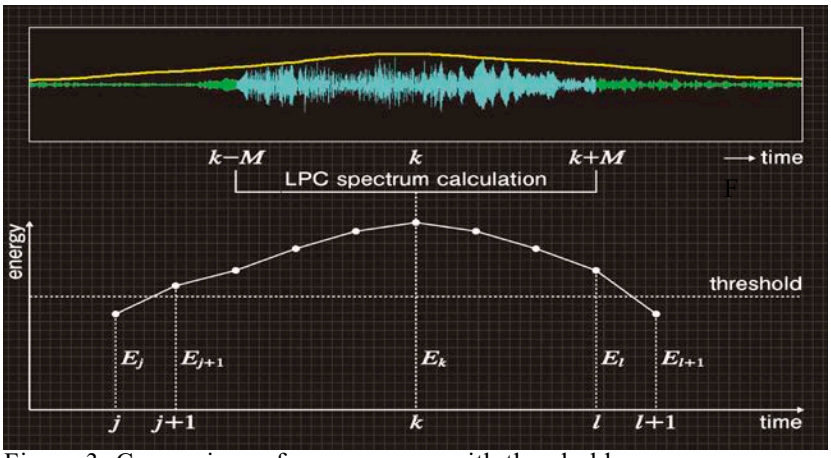

Figure 3: Comparison of energy curve with threshold.

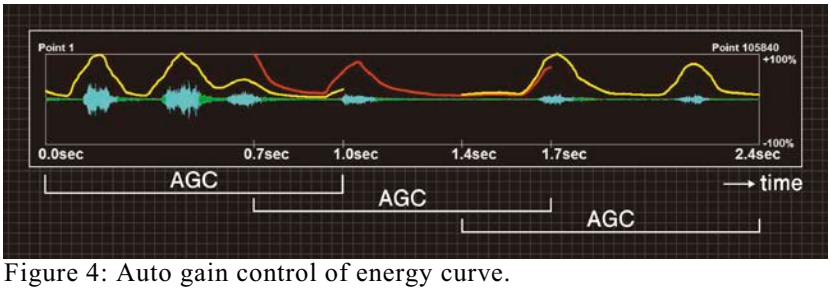

Figure 4: Auto gain control of energy curve.

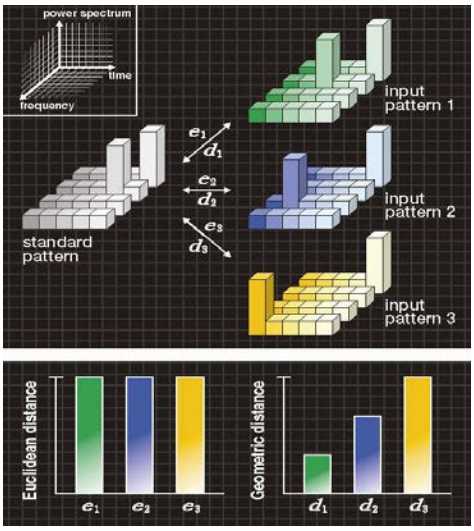

Figure 5: Typical example of "difference".

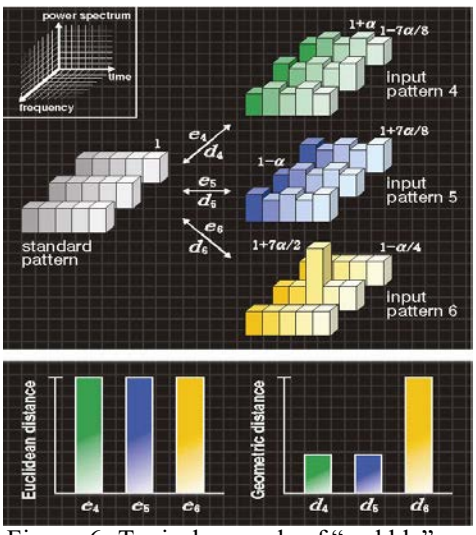

Figure 6: Typical example of "wobble".

5 The similarity scale

Human beings, dogs, cats, and other such animals have "the sense of similarity" in hearing and sight. To realize "the sense of similarity" using an algorithm called "similarity scale" is an important subject for developing computer intelligence. In sound recognition, a known spectrum stored in PC memory is called here the "standard pattern", and a comparison spectrum is called the "input pattern". The degree of likeness between the standard pattern and the input pattern is evaluated using a similarity scale. If the similarity of the standard and input patterns is close, then those two patterns are considered to be in the same category and the input pattern is recognized. The similarity is often measured as a "distance" between the two patterns. Conventionally, the similarity scales known as the Euclidean distance and cosine similarity have been widely used. Section 5.1 describes the shortcomings that are found in the conventional similarity scales. Furthermore, Sections 5.2 describes a new similarity scale called the "Geometric Distance" for improving the shortcomings.

5.1 Conventional similarity scale

Conventional similarity scales the Euclidean distance and the cosine similarity compare the patterns using one-to-one mapping. The result of the one-to-one mapping is that input patterns with different shapes may have the same distance from the standard pattern when the spectra have the "difference" and "wobble". This section describes the shortcomings that are found in the conventional similarity scales using spectrograms (time-frequency-power).

The upper diagram of Figure 5 shows an example of the "difference" where the standard pattern has two peaks in the spectrogram, and input patterns 1, 2, and 3 have a different position on the first peak. Note that both the standard and input patterns have the same volume. As shown in the bar graph at the bottom left of Figure 5, the Euclidean distances and cosine similarities e_1, e_2, and e_3 have the relationship of $e_1=e_2=e_3$ between the standard pattern and each of the input patterns 1, 2, and 3. Therefore, input patterns 1, 2, and 3 cannot be distinguished.

The upper diagram of Figure 6 shows an example of the "wobble" where the standard pattern has a flat spectrogram, input patterns 4 and 5 have the "wobble" on the flat spectrogram, and input pattern 6 has a single peak. However, each pattern is assumed to have variable α in the relationship shown in Figure 6. Therefore, the standard and input patterns always have the same volume. As shown in the bar graph at the bottom left of Figure 6, the Euclidean distances and cosine similarities e_4, e_5, and e_6 have the relationship of $e_4=e_5=e_6$ between the standard pattern and each of the input patterns 4, 5, and 6. Therefore, input patterns 4, 5, and 6 cannot be distinguished.

5.2 New similarity scale

As an improvement, we have developed a new similarity scale called the "Geometric Distance (GD)" [4,5]. A similarity scale is a concept that should intuitively concur with the human concept of similarity in hearing and sight. First we need to develop a mathematical model for the similarity scale so that we can perform numerical processing by computation. In the GD, a mathematical model of the similarity scale is proposed to improve the shortcomings that are found in the Euclidean distance, cosine similarity and others. A mathematical model incorporating the following two characteristics is used:

<1>The distance metric must show good immunity to noise.

<2>The distance metric must increase monotonically when a difference increases between peaks of the standard and input patterns.

The bar graphs at the bottom right of Figures 5 and 6 express the mathematical model diagrammatically. Following on from above, a new algorithm based on one-to-many point mapping is proposed to realize the mathematical model. In this section, we explain the algorithm using the spectrum patterns (frequency-power).

Figure 7 shows the standard and input patterns that have been created using the momentary power spectrum (frequency-power) of standard and input sounds. Figures 8(a)-(e) respectively show typical examples of the standard and input patterns. Note that the power spectrum is generated from the output of a filter bank with the m frequency bands.

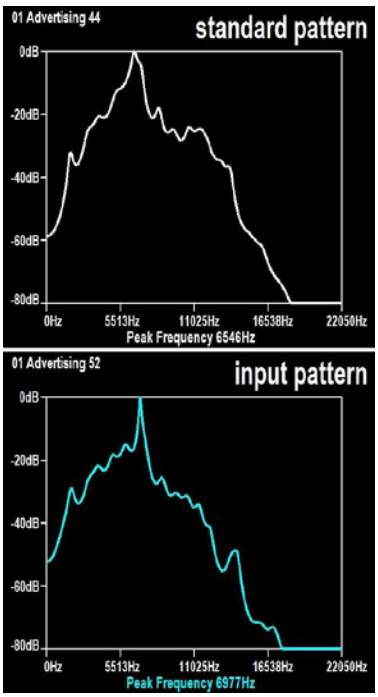

Figure 7: Standard and input patterns.

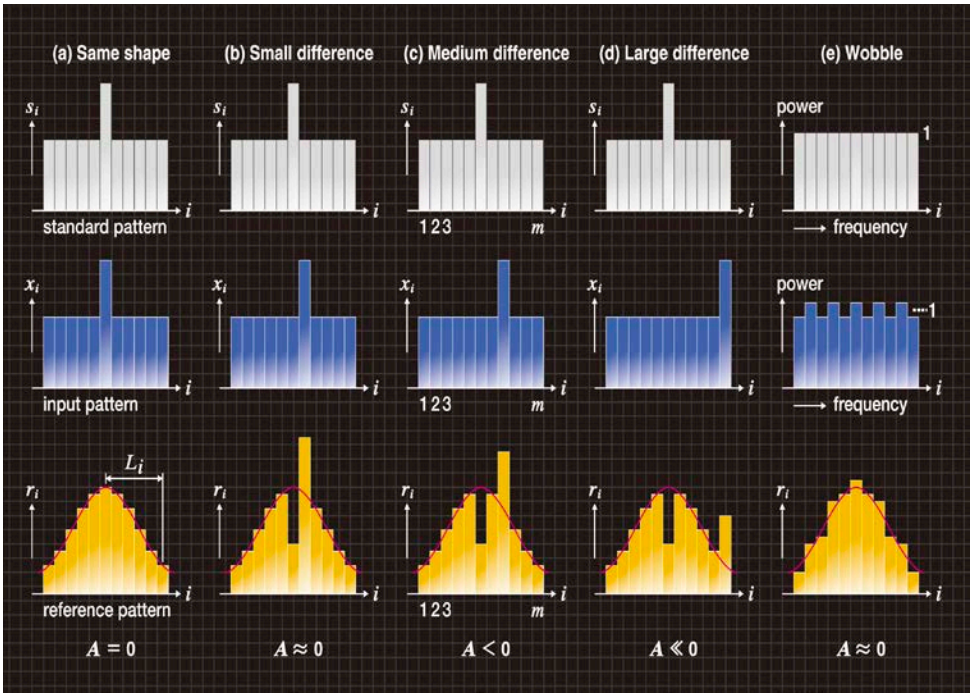

Figure 8: Shape change of reference patterns.

The i-th power spectrum values (where, $i = 1, 2, \ldots, m$) of the standard and input sounds are divided by their total energy, so that normalized power spectra s_i and x_i have been calculated, respectively. At this moment, the standard and input patterns have the same area size. Moreover, Figures 8(a)-(e) respectively show reference patterns that have the initial shape r_i of a normal distribution.

With the GD algorithm, a difference in shapes between standard and input patterns is replaced by the shape change of the reference pattern using the following equation.

$$r_i \longleftarrow r_i + (x_i - s_i) \qquad (i = 1, 2, 3, \cdots, m) \tag{3}$$

Next, we explain Eq. (3) using Figures 8(a)-(e).

● Figure 8(a) gives an example of the case where the standard and input patterns have the same shape. Because values r_i of Eq. (3) do not change during this time, the reference pattern shown in Figure 8(a) does not change in the shape from the normal distribution.

● Figures 8(b)–(d) respectively show examples exhibiting a small, medium, and large "difference" of peaks between the standard and input patterns. If Eq. (3) is represented by the shapes, as shown in Figures 8(b)–(d), value r_i decreases at peak position i of each standard pattern. At the same time, value r_i increases at peak position i of each input pattern.

● Figure 8(e) typically shows the standard pattern having a flat shape and the input pattern where a "wobble" occurs in the flat shape. Because values r_i increase and decrease alternatively in Eq. (3) during this time, the reference pattern shown in Figure 8(e) has a small shape change from the normal distribution.

For the reference pattern whose shape has changed by Eq. (3), the magnitude of shape change is numerically evaluated as the variable of moment ratio. The moment ratio of the reference pattern can be calculated using the following equation.

$$A = \frac{\left\{\sum_{i=1}^{m} r_i\right\} \cdot \left\{\sum_{i=1}^{m} (L_i)^4 \cdot r_i\right\}}{\left\{\sum_{i=1}^{m} (L_i)^2 \cdot r_i\right\}^2} - 3 \tag{4}$$

Where, L_i $(i = 1, 2, \ldots, m)$ is a deviation from the centre axis of the normal distribution as shown in the reference pattern of Figure 8(a). The moment ratio A is derived from the kurtosis from a statistical analysis. If the shape of the reference pattern follows the normal distribution, then $A = 0$. If it has peakedness relative to the normal distribution, then $A > 0$. Alternatively, if it has flatness relative to the normal distribution, then $A < 0$. Figures 8(a)–(e) show how A varies with r_i.

● In Figure 8(a), the values r_i do not change. The moment ratio becomes $A = 0$.

● In Figure 8(b), the position i of the decreased r_i and that of the increased r_i are close. Because the effect of an increase and a decrease is cancelled out, the moment ratio becomes $A \approx 0$.

● In Figure 8(d), because the shape of the reference pattern has flatness relative to the normal distribution, the moment ratio becomes $A \ll 0$.

● In Figure 8(c), because the shape of the reference pattern is an intermediate state between (b) and (d), the moment ratio becomes $A < 0$.

● In Figure 8(e), the reference pattern has a small shape change from the normal distribution, and the moment ratio becomes $A \approx 0$.

From Figures 8(a)–(d), we can understand that the value $|A|$ increases monotonically according to the increase of the "difference" between peaks of the standard and input patterns. Also, from Figure 8(e), it is clear that $A \approx 0$ for the "wobble".

In Figures 8(a)–(e), we have explained the case where the standard and input patterns have a single peak respectively. Next, in Figures 9(a) and (b), we explain the case where they have two peaks respectively. Figure 9(a) shows a typical example of the reference pattern that has a large variance value of the normal distribution. Because the positions of two bars (i) are symmetrical about the centre axis of the normal distribution, the effect of an increase and a decrease is cancelled out. Similarly, the effect of two bars (ii) is cancelled out. The moment ratio becomes $A_j = 0$. As an improvement, Figure 9(b) shows a typical example of the two reference patterns that have a small variance value of the normal distribution. In Figure 9(b), the moment ratio becomes $A_4 < 0$ and $A_j < 0$ in the same way as Figure 8(c).

If the small variance value of the normal distribution is used, then we need to move the reference pattern so that it covers the standard and input patterns. Therefore, as shown in Figure 10, we determine the amount of moment ratio A_j for each j in the case where the centre axis of the normal distribution moves to any component position j (where, $j = 1, 2, \ldots, m$) of the standard and input patterns. Using the m parts of the moment ratios A_j that we have obtained in Figure 10, we can calculate the difference in shapes between standard and input patterns by the following equation and we define it as the "Geometric Distance d".

$$d = \sqrt{\sum_{j=1}^{m} (A_j)^2} \tag{5}$$

In this method, when a "difference" occurs between peaks of the standard and input patterns with a "wobble" due to noise, the "wobble" is absorbed and the distance metric increases monotonically according to the increase of the "difference". From the above description, we could verify that the GD algorithm matches the characteristics <1> and <2> of the mathematical model.

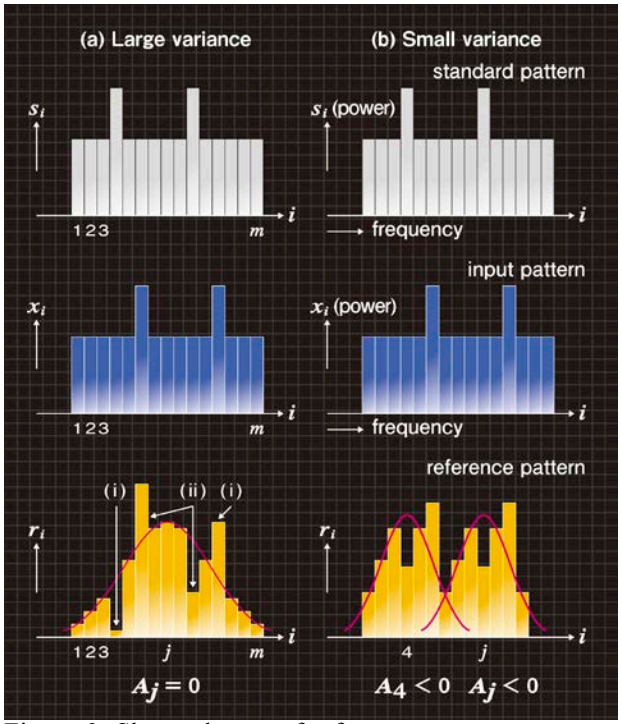

Figure 9: Shape change of reference patterns.

6 Optimization of geometric distance

The proposed technique replaces the amount of "difference" between peaks of the standard and input patterns by the shape change of normal distribution and detects it. In such a case, as shown in Figures 9(a) and (b), it is important to optimize the shape (variance σ^2) of normal distribution that covers the standard and input patterns. In this section, we describe a new optimization method of the Geometric Distance.

In the reference patterns shown in Figures 10(c)-(f), the "bright" bar graph corresponds to the component number i of the input pattern and, therefore, its value changes according to the "wobble" of the input pattern. However, the "dark" bar graph does not correspond to it and its value does not change. The sensitivity to the "wobble" in the reference patterns is equated regardless of the movement position of the normal distribution. Therefore, we set value "ω" so that the number of bright bar graphs is equated in all the reference patterns. In Figures 10(c)-(f), for example, each reference pattern consists of 7 bright bar graphs (ω=7).

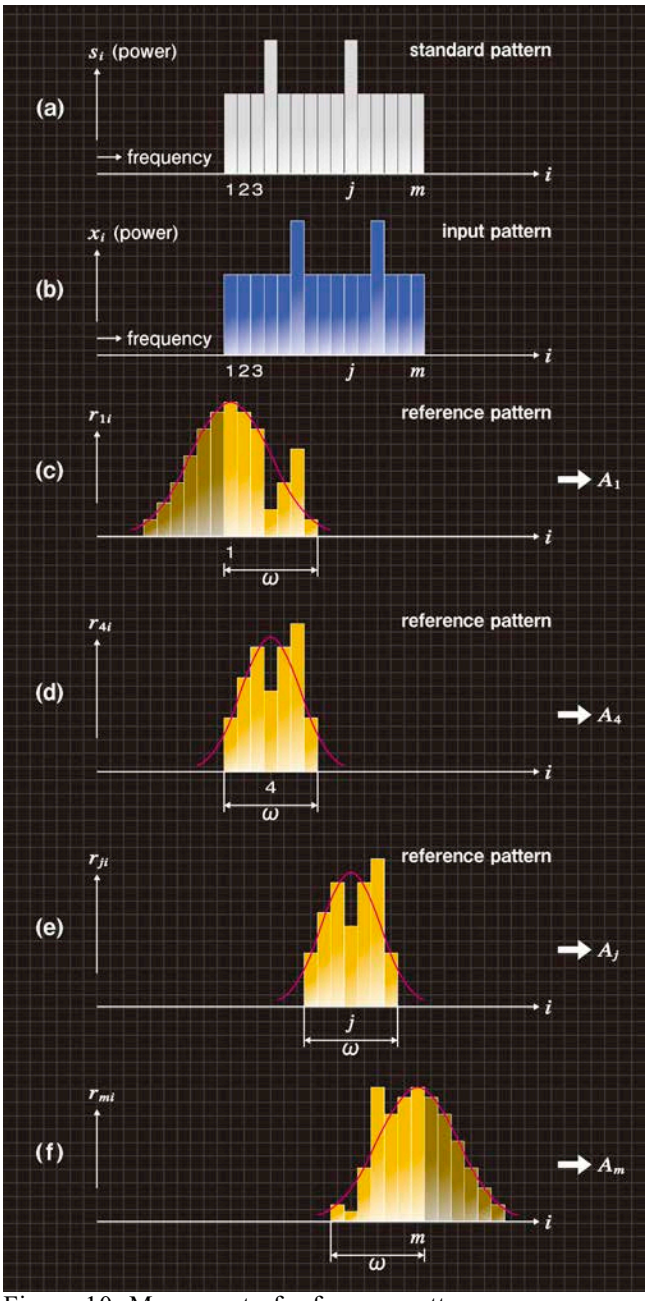

Figure 10: Movement of reference pattern.

For example, we suppose that 10 spectra are extracted from 10 vocalisations of a target bird and that also 10 spectra are extracted from the "other" 10 vocalisations of the same target bird. Then, we calculate the GD value d between their spectra using an arbitrary value ω. The 10×10 pieces of the GD values are calculated for each combination of the 10 spectra obtained from the vocalisations of the target bird and the 10 spectra obtained from the "other" vocalisations of the same target bird. Furthermore, we perform logarithmic transformation for the 10×10 pieces of the GD values, so that the GD values meet the normal distribution. Note that N_1=10×10, $\bar{x}_1$, and s_1^2 are the sample size, sample mean, and sample variance of the GD values, respectively. Figure 11 shows a typical example of the distribution of the GD values that are calculated using the arbitrary value ω. Similarly, 10 spectra are extracted from 10 vocalisations of one target bird and 10 spectra are extracted from 10 vocalisations of non-target birds. Then, the sample size N_2=10×10, sample mean $\bar{x}_2$, and sample variance s_2^2 of the GD values are calculated. In this paper, we have adopted a statistic T of "Welch's T-test" as an objective function in order to measure the degree of separation between both the target and non-target birds' vocalisations. The statistic T can be calculated using the following equation.

$$T = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \tag{6}$$

The LPC spectrum shown in Figure 7 has 256 bars. In order to determine the optimum value ω, we have scanned the value ω by 2 from 1 to 239 and calculated the GD values. Figure 12 shows the calculated relationship between the value ω and the objective function T. From Figure 12, it is discovered that the objective function T becomes maximum if ω=96. Thus, we determine ω=96 as the optimum value and use it in the following evaluation experiment. Note that, instead of Eq. (6), we can also use

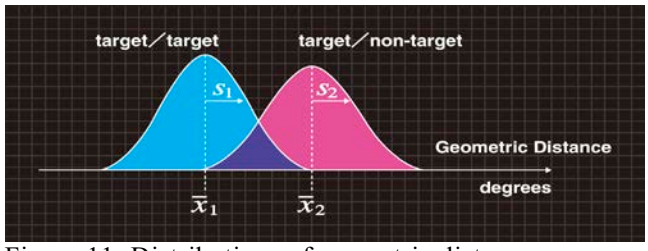

Figure 11: Distributions of geometric distances.

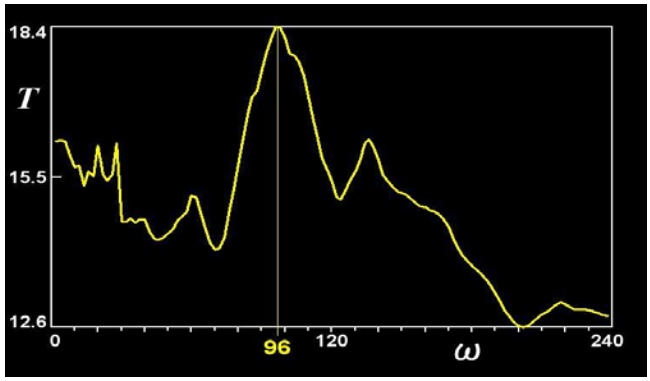

Figure 12: Objective function and optimum value ω.

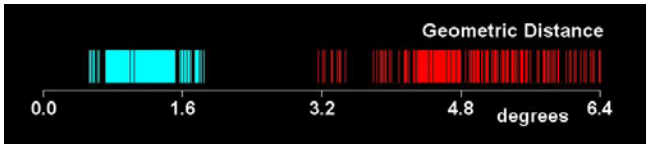

Figure 13: Distributions with optimum value ω=96.

an area of a critical region that is derived from Welch's *T*-test as the objective function. However, the statistic *T* of Eq. (6) usually has a large value in the case of bird vocalisations, while the area of a critical region is a quite small value. Therefore, we use Eq. (6) as the objective function.

We have performed the evaluation experiment using the optimum value ω=96 determined from Figure 12. Figure 13 shows the distributions of the GD values with the value ω=96. In Figure 13, the GD values between the target and non-target birds' vocalisations are shown by red lines, and the GD values between one target bird's vocalisations and the other target bird's vocalisations are shown by blue lines. From Figure 13, it is clear that the blue lines are separated from the red lines completely and definitively.

7 The spectrogram of bird sounds

The upper diagram of Figure 14 shows a spectrogram (time-frequency-power) extracted from the vocalisation of the Coxen's Fig-Parrot. As shown at the bottom of Figure 14, the waveform has been segmented with 11.4 msec frame width and 0.68 msec frame period, and the LPC spectrum has been calculated in each frame. Next, the spectrogram has been coloured according to logarithmic power of the LPC spectrum. We have set the analysis conditions of the bird vocalisation with a 44.1kHz sampling frequency, 16 bit quantization, 16 order LPC, 1184Hz to 12209Hz frequency range, 5.38Hz frequency resolution, and 0dB to −80dB logarithmic power spectrum.

If we analyze transient signals such as bird vocalisations, we then need to set the short frame width as shown at the bottom of Figure 14. The LPC spectrum analysis is suitable for such transient signals. We have developed the 64 bit software so that it extracts spectrograms from the bird vocalisations using the LPC and matches images of the spectrogram using the Two-dimensional GD [3].

8 The biological significance

The acoustic constructs of nonhuman animals are under regular study for their importance in understanding vocal learning, sound perception and processing, and evolutionary patterns and processes. The spectrograph is effective in documenting sounds whose frequency and speed fall outside of human sensory norms. It is employed to make taxonomic identifications, distinguish song types, and create genetic profiles. American ornithologist Donald Kroodsma exemplifies the reliance on the visual aspect of spectrographs: "You must have well-trained ears," people often say to me.... "No," I reply, they're actually pretty pathetic, and I have no musical ability whatsoever. But, like most of us, I have well-trained eyes, and it is with my eyes that I hear" [6]. Given biologists' strong reliance on the spectrograph, biologists stand to benefit from our automatic recognition system for bird vocalisations and our proposed new optimization method in these areas.

9 Conclusions and future work

We have introduced an automatic recognition system for bird vocalisations and proposed a new optimization method of the Geometric Distance. The software segments a waveform of bird vocalisations from up to three hours of continuous recordings and extracts the sound spectrum pattern from the waveform using the LPC spectrum analysis. Next, the software matches the spectrum patterns using the GD and recognize the bird vocalisation. We have performed the optimization experiment and verified the effectiveness of the method.

In our future work, we will continue to develop the 64 bit recognition system using the Two-dimensional Geometric Distance. Furthermore, we will develop an optimization method of the Two-dimensional GD and will verify the effectiveness of the method.

References

[1] N. Boucher, A. Burbidge, M. Jinnai, "Computer recognition of sounds that have never been heard before" *Australian Institute of Physics 18th National Congress*, 211, (2008) http://www.soundid.net/

[2] N. Boucher, M. Jinnai, H. Taylor, "A new and improved four-dimensional spectrogram" *Australian Institute of Physics 19th National Congress*, 210, (2010) http://www.soundid.net/

[3] M. Jinnai, N. Boucher, J. Robertson, S. Kleindorfer, "Design consideration in an automatic classification system for bird vocalisations using the two-dimensional geometric distance and cluster analysis", *20th International Congress on Acoustics*, 130, (2010) http://www.soundid.net/

[4] M. Jinnai, S. Tsuge, S. Kuroiwa, F. Ren, M. Fukumi, "New similarity scale to measure the difference in like patterns with noise", *International Journal of Advanced Intelligence*, Volume 1, Number 1, pp. 59-88 (2009) http://aia-i.com/ijai/contents.html

[5] M. Jinnai, S. Tsuge, S. Kuroiwa, M. Fukumi, "A new geometric distance method to remove pseudo difference in shapes", *International Journal of Advanced Intelligence*, Volume 2, Number 1, pp. 119-144 (2010) http://aia-i.com/ijai/contents.htm

[6] D. Kroodsma, *The Singing Life of Birds*. New York: Houghton Mifflin, pp. 1 (2005).

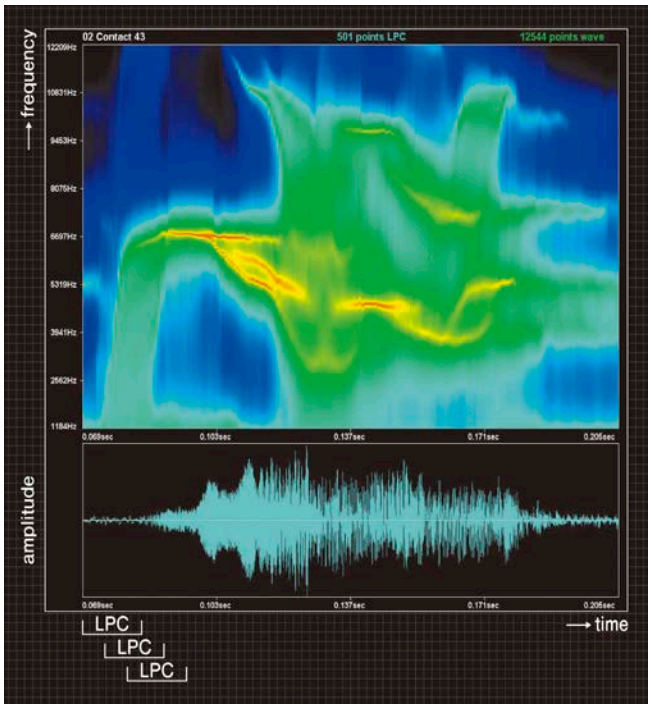

Figure 14: Spectrogram of bird vocalisation.